A Web-accessible content-based cervicographic image retrieval system

Zhiyun Xue*[a], L. Rodney Long[a], Sameer Antani[a], Jose Jeronimo[b], George R. Thoma[a]
[a]National Library of Medicine, NIH, Bethesda, MD, USA 20894;
[b]National Cancer Institute, NIH, Bethesda, MD, USA 20894

ABSTRACT

Content-based image retrieval (CBIR) is the process of retrieving images by directly using image visual characteristics. In this paper, we present a prototype system implemented for CBIR for a uterine cervix image (cervigram) database. This cervigram database is a part of data collected in a multi-year longitudinal effort by the National Cancer Institute (NCI), and archived by the National Library of Medicine (NLM), for the study of the origins of, and factors related to, cervical precancer/cancer. Users may access the system with any Web browser. The system is built with a distributed architecture which is modular and expandable; the user interface is decoupled from the core indexing and retrieving algorithms, and uses open communication standards and open source software. The system tries to bridge the gap between a user's semantic understanding and image feature representation, by incorporating the user's knowledge. Given a user-specified query region, the system returns the most similar regions from the database, with respect to attributes of color, texture, and size. Experimental evaluation of the retrieval performance of the system on "ground-truth" test data illustrates its feasibility to serve as a possible research tool to aid the study of the visual characteristics of cervical neoplasia.

Keywords: content-based image retrieval, Web-based medical image system, uterine cervix cancer

1. INTRODUCTION

With the fast growing volume of digitized medical images used for clinical diagnosis and treatment, the research of content-based image retrieval (CBIR) for medical applications has become increasingly active in recent years [1-7]. In contrast to text-based image retrieval that uses textual language to describe the image content and consequently has significant limitations since image data cannot be fully described texturally, CBIR directly utilizes visual characteristics, such as color, texture, and shape, to represent image content and to retrieve images from image databases which are visually similar to a given query image. As indicated in [8], one big challenge of CBIR is to bridge the semantic gap between the low-level features extracted automatically from images by machines and the high-level concepts/interpretation of humans. Although the semantic gap may be wide in a broad image domain, it is usually smaller in a narrow domain of medical images. Therefore, by incorporating experts' explicit domain knowledge, CBIR for medical databases has the potential to assist clinical decision-making, research, and training, in addition to aiding medical image data management and analysis.

Our medical image database contains cervicographic images (also called cervigrams) and was created by the collaborative efforts of the National Cancer Institute (NCI) and the National Library of Medicine (NLM) for the study of uterine cervix cancer. Cervical cancer is the second most common cancer affecting women worldwide and the first in many developing countries. This cancer is closely related to the chronic infection of certain types of Human Papillomavirus (HPV). To visually screen for pre-invasive cervical lesion or for cancer, one cost-effective method is cervicography. Cervicographic screening is based on the acetowhitening phenomenon: HPV-infected abnormal tissue often turns white after being treated with 3-5% acetic acid. A cervigram is a 35-mm photograph of the cervix taken at approximately one minute after acetic acid exposure. For this photography, a specially designed fixed-focus optical camera equipped with a macro lens and ring flash is used. Our cervigram database contains approximately 100,000 cervigrams taken during two major projects in cervical cancer carried out by NCI to study the natural history of HPV infection and cervical neoplasia, the *Guanacaste* and *ALTS* projects. In addition to cervigrams, correlated clinical,

*xuez@mail.nih.gov; phone 1 301 435-3260

Medical Imaging 2008: PACS and Imaging Informatics, edited by Katherine P. Andriole, Khan M. Siddiqui,
Proc. of SPIE Vol. 6919, 691907, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.769440

cytologic and molecular information were also collected. To store and provide access to this large amount of information, NLM has been developing the Multimedia Database Tool (MDT), which has the capability to query on any of the text data fields, such as age, Pap smear results, or HPV status, to retrieve clinical records from the database, as well as associated images. Complementary to the text-query based MDT, CBIR queries may become significant modes of accessing data from future databases that combine images and text.

In this paper, we describe our Web-accessible CBIR prototype system which operates on a subset of the cervigram database described above. To use our system, the user creates a query region by marking a region of interest on an image through the graphical user interface. The system calculates the feature vector of the query region for the specified features and compares it with the pre-computed feature vectors of regions stored in the database. The returned regions are ranked by the degree of similarity to the query feature vector and presented on a multi-image display, along with associated text information. The (visual) features used are color, texture and size. The shape feature is significantly less important for identifying or distinguishing regions in our application since these tissue types do not have specific shape except for os[1] regions which are somewhat elliptic. In order to facilitate medical experts located in geographically different places to help us evaluate the prototype system, as well as to allow the final system to be accessed remotely for either diagnosis or learning in the future, the system is designed and implemented using a distributed client/server framework. It has the capability of accommodating large amounts of imaging data.

The rest of this paper is organized as follows. A brief description of the image data used by the prototype system is given in Section 2. The feature representations, feature weighting and combination, and similarity measures employed in the CBIR system are presented in Section 3. Section 4 describes the system architecture and the Web-interface. The evaluation of the system is presented in Section 5, followed by the conclusion given in Section 6.

2. IMAGE DATA

Figure 1 shows two examples of cervigrams in the database. As shown in both examples, in addition to the cervix region in which we are interested, a cervigram may contain irrelevant information, such as medical instruments, film markup, vaginal walls and other non-cervix anatomy or regions. Inside the cervix region, in addition to the acetowhite region, which is a high-interest biomarker as it may be potentially malignant, other anatomical regions, such as the squamous epithelium (SE), columnar epithelium, cyst, blood, mucous, polyps, cervical os and squamous metaplasia, are also of clinical significance. Therefore, for querying and retrieving the cervigram database, we aim to develop a region-based CBIR system in which the features of the region-of-interest are extracted and compared, in contrast to using global features calculated from the whole image. Global features are suitable for medical image categorization and general stock picture retrieval, while our cervigram application requires local features to capture the pathology that is localized in particular regions within the image. Therefore, automatic image segmentation and region type recognition are initial steps necessary for CBIR. Our ongoing research on this topic has been reported in [11-14]. The prototype CBIR system operates on a subset of the cervigram database in which important regions were manually marked and labeled by dozens of NCI medical experts, with the assistance of our Boundary Marking Tool (BMT). The BMT is a Java application that allows expert colposcopists at geographically-distributed sites to mark the boundaries of regions corresponding to important tissue types and to record the labels of regions as well as detailed feature information [15]. A screen shot of the graphical user interface of the BMT tool showing an expert-annotated cervigram is given in Figure 2.

[1] The os is the opening to the cervix.

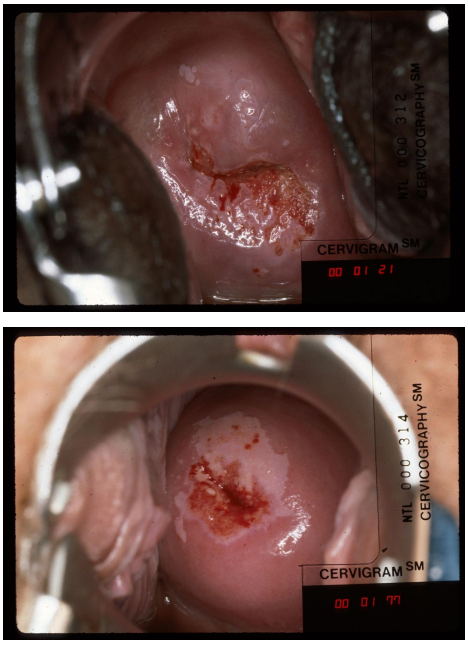


Fig. 1. Cervigrams

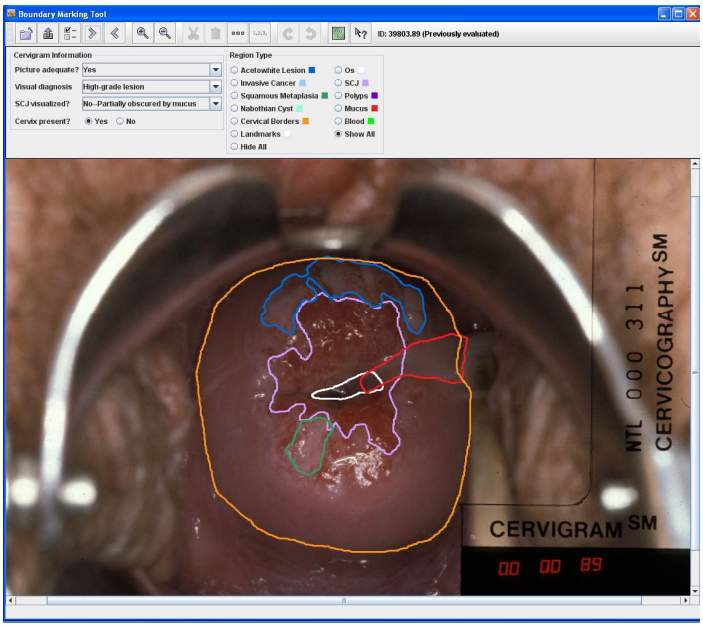


Fig. 2. Tool for manual segmentation and labeling of regions

3. FEATURE REPRESENTATIONS AND COMBINATION

For visual inspection of cervix lesions, the color tone and surface texture are two major features used by physicians to differentiate tissue types and to identify the developing stage of cervical neoplasia. For example, the acetowhite region shows a certain degree of opacity and whiteness. The squamous epithelium tissue is smooth and appears pinkish in color. Metaplastic squamous epithelium often appears mildly white compared to the pink original squamous epithelium. The columnar epithelium is red or orange and often has a grainy and grape-like appearance. Low-grade cervical intraepithelial neoplasia (CIN) is often seen as thin, smooth acetowhite lesions, while high-grade CIN is associated with thick, dense, greyish-white acetowhite areas [16]. Therefore, color and texture features are extracted and used for query indexing and retrieval in the prototype system. Additionally, the size attribute is also used to search regions through the image database, since the size of the lesion is one indicator of the degree of severity.

3.1 Color feature

For quantifying color content in images, histogram-based representations, such as the color histogram, color coherence vector, and color correlogram, are popular approaches that have been employed in various image retrieval applications. They work reasonably well in images whose colors are widely distributed in the whole color space. In our system, the objects of interest are homogeneous regions, in which the color content is narrowly distributed in the color space. Therefore, we use the color moments descriptor, which offers computational simplicity, speedy retrieval, and minimal storage, and does not require the procedure of color quantization which is usually needed by histogram-based color descriptors. For each region, the mean, the standard deviation and the skewness of each color channel are computed. The CIE-Lab color space is used, since it is a perceptually uniform color space.

3.2 Texture feature

To analyze the texture of images in the CBIR system, the features are derived from the wavelet coefficients obtained by applying the Discrete Wavelet Transform (DWT). The DWT requires filtering (a quadrature mirror filter) and subsampling. It recursively decomposes an image into low frequency and high frequency sub-images. Specifically, in the first-level wavelet transform, a pair of wavelet filters including a low-pass filter and a high-pass filter is applied to the image along the rows and columns, and the output images of the filtering are then downsampled by two. This generates four sub-images of the original image: the approximation sub-image, the horizontal detail sub-image, the vertical detail sub-image, and the diagonal detail sub-image. A multi-level wavelet transformation is then obtained by applying the same procedure to the approximation sub-image at subsequent levels. We use a 4-level wavelet transform, which results in 13 sub-images. For each sub-image, the mean and standard deviation of the magnitude of its wavelet coefficients are calculated as the features. It should be noted that, in order to extract the wavelet features of a region (not the whole image), the wavelet transform of the binary mask of the region-of-interest is obtained at the same time. Therefore, for each sub-image, only the coefficients inside the sub-mask are considered. These texture features as well as the color moment features are normalized using Gaussian normalization to approximately equalize their scale.

3.3 Size feature

As previously stated, the camera for taking cervigrams is fixed-focus. Consequently, the distance between the camera and the cervix is expected to be approximately constant; this allows us to obtain comparable pictures of cervices of all patients with respect to the region sizes. In the system, two options are provided for comparing the size of lesions. One is the absolute size (the size of lesion is normalized to the image size). The other is the relative size (the size of lesion is normalized to the size of the cervix if the cervix boundary information is available).

3.4 Feature combination

Sometimes, users are interested in representing and comparing the region content by using multiple features. Therefore, these features need to be combined together to decide the final retrieval results. In our system, the feature classes (color, texture, and size) are combined at the "rank level". That is, for each feature class, the system compares the feature vector of the query region to those of regions in the database, using a pre-determined similarity measure (such as Euclidean distance), and ranks the retrieved regions/images by their relevance score. Then the ranks of the different feature classes are combined into one overall rank, and we use this overall rank to define the overall similarity of database regions.

The degree of importance of different feature classes is allowed to be different across the classes. Our system implements this capability by using a weighted linear method to combine the feature classes. Then, overall similarity is calculated as

$$\bar{r} = \sum_{i=1}^{n} \omega_i \bar{r}_i \tag{1}$$

where ω_i and $\bar{r}_i$ are the weight and rank vector for the *i*-th feature class. The set of weights of the feature classes ($\omega_i, i = 1,...n$) is directly specified by the user.

4. SYSTEM ARCHITECTURE

To provide a remote content-based retrieval service for the image database, our system uses the distributed architecture of the Spine Pathology & Image Retrieval System (SPIRS), a Web-accessible system developed by our group for performing image retrieval on a database of digitized spine x-rays using the morphological shape of the vertebral body [17]. This system employs open communication standards and open source software, and decouples the user interface from the core indexing and retrieving algorithms. As shown in Figure 3, the system architecture consists of four modules:

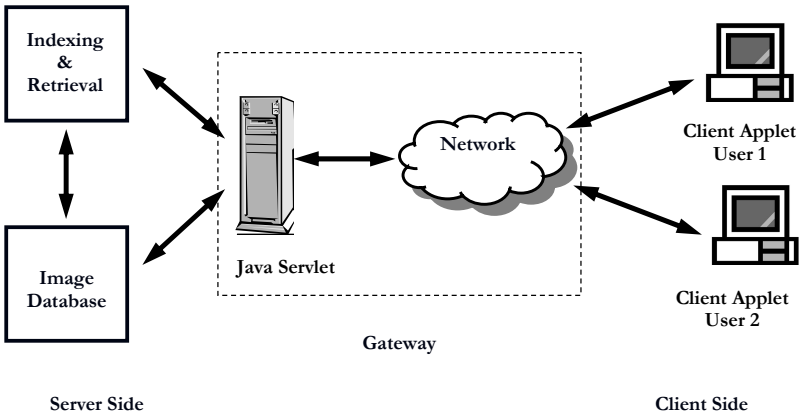


Fig. 3. The distributed architecture of the system

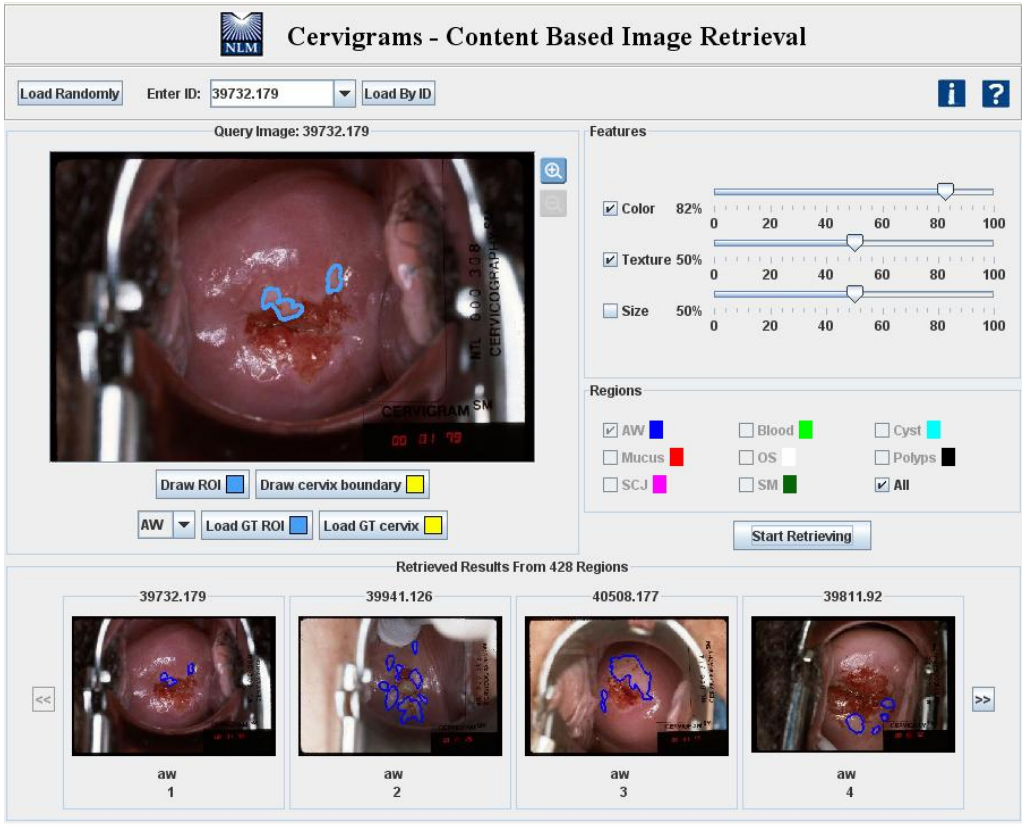


Fig. 4. The client applet GUI

1) A client that provides a graphical user interface

The client is a Java applet that runs in a Web browser. It contains four panels for selecting an image, creating a query, specifying system parameters, and displaying retrieved results, respectively, as shown in Figure 4. Query images can be selected from the image repository randomly (the system will choose one for the user), or by user specification. A query region can then be created by manually delineating the region-of-interest (ROI) on the query image using the mouse, or

by loading the ROI from the database of regions pre-marked by medical experts. Image zoom is provided. After selecting the visual features that he or she is interested in, specifying the weights if multiple features are selected, and choosing the types of regions to be searched, the user can initiate the retrieval process with the "Start" button. Images containing the regions which are found by the system to be similar to the user's query region (based on the selected features and specified weights) will be presented in the retrieval panel in descending order of relevance. The corresponding image database tag, region labels and rank number are also displayed. A magnified view of each returned image is available to the user by clicking within the pixel area of any returned image. Help information on the functionality of each component in the GUI is also provided.

2) A gateway which uses a Java servlet for Web client-server communication

The system uses a Java servlet to manage multiple requests and responses between the client applet and the components on the server side. The servlet (1) sends the query information collected by the client applet to the indexing and retrieval server, and (2) reformats and returns to the client the query results it obtains from the indexing and retrieval server. (The client then displays the query results.). It also performs user authentication functions. The servlet communicates with a client applet using the https protocol and communicates with the indexing and retrieval server through a TCP socket interface.

3) The indexing and retrieval server which extracts features and computes similarity

The indexing and retrieval server calculates the feature signature of the query region and compares it with pre-computed signatures extracted from each region in the database. The retrieval process is implemented in a modular way to facilitate incorporation of new feature representation methods and similarity measures as they become available. Currently, we have implemented various popular algorithms for representing color and texture properties. For color, four additional color descriptors besides color moments have been implemented. They are the color histogram, color coherence vectors, color correlograms, and dominant colors. To extract the texture feature, in additional to the DWT-based approach, we have implemented four methods which are based on the Gabor filter, the Log-Gabor filter, the gray-level co-occurrence matrix, and the histogram of edge directions, respectively. For similarity measures, we also have implemented multiple distance-based measures, including Minkowski-form distance, histogram intersection, Jeffrey divergence, quadratic-form distance, and Earth mover's distance.

4) The image-text database

The complete database consists of the approximately 100,000 cervigrams and other uterine cervix data collected by NCI. It is intended to be used for the study of the relationship between the natural history of HPV infection and cervical cancer as well as for education and training of professionals for prevention and treatment of this high incidence disease. In our system, a subset of this database is accessed by both the indexing and retrieval server and by the gateway. The images used for testing the prototype CBIR system are "ground truth" images with regions marked and labeled by multiple experts.

5. EVALUATION AND DISCUSSION

Precision is one of the most popular evaluation measures used in the domains of information retrieval, and we have adopted it for CBIR evaluation. It is defined as,

$$precision = \frac{number\ of\ retrieved\ regions\ which\ are\ relevevant}{number\ of\ regions\ retrieved} \tag{2}$$

We note that the CBIR system is for finding images with regions similar to the specified regions on a query image where the retrieval results may be tuned by the individual user through the specification of the particular feature classes and relative weights to be used for the retrieval. Therefore, the relevance of the results is extremely subjective, and the ultimate evaluation of the system should take into account the user requirements by setting up comprehensive subjective experiments. This is time consuming, difficult, and expensive to accomplish, especially for medical applications. As a workaround approach to evaluation, we take a different track: we judge the retrieval performance by classification accuracy, as follows: in our ground truth dataset, each region has a region-type label assigned by experts; we test the system by using a particular region type as query, and evaluate the returned results for the number of matching region

types. In tests done to date, we presented each of the expert-marked region types in turn to the system as a query and calculated the fraction of the top (five) returned images that had the same region type as the query. The accuracy measure is defined below,

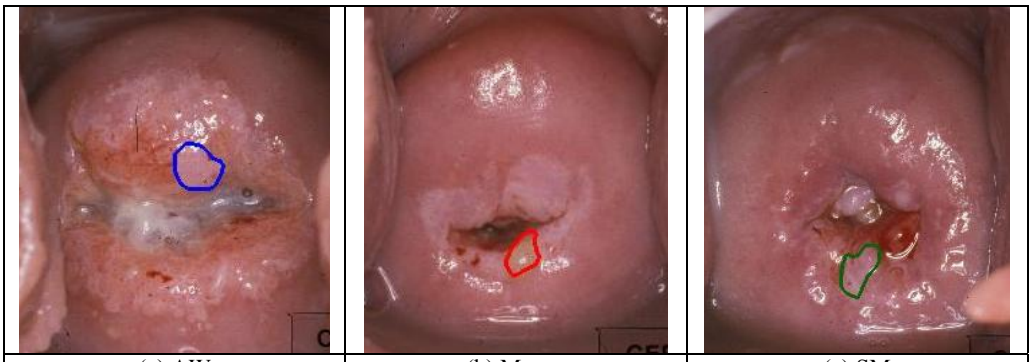
$$accuracy = \frac{number\ of\ retrieved\ regions\ which\ are\ classified\ correctly}{number\ of\ regions\ retrieved} \tag{3}$$

The ground truth data set we used for test has 120 images with a total of 422 tissue regions marked by medical experts. There are six types of tissues: acetowhite (AW), blood, mucus, os, columnar epithelium (CE), and squamous metaplasia (SM). Table 1 gives the results of the above evaluation of retrieval of matching region types. The table lists the average accuracy for retrieving of each region type. For each region type, a query was made using each of the features: color, texture, and size. For the query by color feature, color was weighted 100%, texture 0%, and size 0%; the queries by texture and size were weighted analogously. The first column of Table 1 shows the number of cases for each tissue type. Table 1 indicates that the importance of each feature class for differentiation is different for each tissue type. For example, color is crucial to identify the blood region while the size of os is relatively more consistent. In Table 1, the accuracy scores for tissue types Mucus and SM are relatively low. This can be explained by the close visual similarity between them and AW as shown in Figure 5 and the large appearance variations across images due to irregular illumination and different acquisition conditions.

Comparing equation 2 and 3, the accuracy measure is equivalent to the precision measure if the definition of 'being relevant' is the same as that of 'being classified correctly'. Therefore, this evaluation approach is based on the assumption that the regions of the same label have the same characteristics with respect to the visual features in the user's mind, while the regions of different labels do not. This assumption may not be totally true for some cases since the labels in the ground truth data annotated by experts are based on the integration of all characteristics while the user may only be interested in only one aspect of the characteristics. For example, if the user is interested in retrieving regions with respect to color similarity, the two aspects of the assumption become: 1) those returned regions having the same label as the query region have similar color to the query; and 2) those returned regions not having the same label as the query region do not have similar color to the query. The first assumption is quite true for most region types, since they have relatively consistent color, while the second assumption is not quite true for some cases as shown in Figure 5, where the colors of three regions labeled as AW, mucus and SM respectively, are very close to each other. In fact, sometimes the difference between those regions is very subtle, and it requires considerable skill for medical experts to differentiate them. Given one query region such as the AW region in Figure 5 and selecting the color feature, the top returned regions retrieved from the database may have labels of mucus or SM, although their colors are close to the color of the query region. Therefore, the real average precision score should be higher than the average accuracy score given in Table 1.

Table 1. The average accuracy for each feature class (Quantity is also number of queries)

Type	Quantity	Color	Texture	Size
AW	76	**51.3%**	46.3%	41.8%
Blood	51	**65.1%**	23.9%	13.7%
Mucus	40	**14.5%**	4.5%	7.0%
Os	120	49.3%	35.0%	**55.8%**
CE	77	48.8%	**49.4%**	40.3%
SM	58	**35.5%**	19.3%	19.3%
Total	422	**46.3%**	33.3%	35.7%



(a) AW (b) Mucus (c) SM

Fig. 5. Examples of AW, Mucus and SM

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a prototype CBIR system that searches and retrieves similar images from a uterine cervix image database based on the properties of color, texture and size. It uses a simple but effective descriptor, color moments, to represent the color attribute. The texture information is represented by features extracted from coefficients after images are decomposed using the discrete wavelet transform. The system combines the image characterization obtained from automatic image processing and from user knowledge (for the selection of the region-of-interest and its attributes) to bridge the gap between the user's understanding and the region representation in the database. The system has a distributed architecture with the advantages of simplicity, extensibility, flexibility, and security. It can be accessed using a Web browser. Preliminary empirical assessment of the system has demonstrated its potential for being used as a tool to assist the study of visual precursors of cervical cancer.

We are working on several aspects to further improve the retrieval performance. Currently, the color feature is calculated from the original image. However, due to the uneven surface of the cervix and different acquisition conditions, illumination inhomogeneity exists in every cervigram and is variable across images in the database. The varying illumination conditions could affect the accuracy of comparing images across the database. We will embed an illumination correction and normalization method into our CBIR system and extract features using illumination corrected images instead of original images to test if the retrieval performance is improved. In addition to color, texture, and size features, the spatial property is also used by physicians to characterize the visual appearance of lesions. For example, the physicians use the term "at 12 o'clock position" or "in the right upper quadrant" to describe the location of a lesion. We will add the function of retrieving similar regions with respect to their location features into the system. Further evaluation of the system's retrieval performance with the collaboration of medical experts from NCI as well as on a larger dataset is also planned.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

1. H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions", International Journal of Medical Informatics, Vol. 73, No. 1, pp. 1-23, 2004.

2. M. O. Güld, C. Thies, B. Fischer, T. M. Lehmann, "A generic concept for the implementation of medical image retrieval systems", International Journal of Medical Informatics, Vol. 76, pp.252-259, 2007.
3. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C.E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment", Radiology, Vol. 228, No.1, pp. 265-270, 2003.
4. S. Antani, J. Cheng, J. Long, L. R. Long, G. R. Thoma, "Medical validation and CBIR of spine X-ray images over the internet", Proceedings of SPIE Electronic Imaging Science and Technology, Vol. 6061 pp. 60610J-1-9, 2006.
5. E. G. M. Petrakis, "Content-based retrieval of medical images", International Journal of Computer Research, Vol. 11, No. 2, pp. 171-182, 2002.
6. L. R. Long, S. Antani, G. R. Thoma, "Image informatics at a national research center", Computerized Medical Imaging and Graphics, Vol. 29, pp. 171-193, February 2005.
7. M. Siadat, H. Soltanian-Zadeh, F. Fotouhi, K. Elisevich, "Content-based image database system for epilepsy", Computer Methods and Programs in Biomedicine, Vol. 79, No. 3, pp. 209-226, 2005.
8. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1379, December 2000.
9. R. Herrero, M. H. Schiffman, C. Bratti, A. Hildesheim, I. Balmaceda, M. E. Sherman, "Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste project", Rev Panam Salud Publica, No. 1, pp. 362-375, 1997.
10. M. Bopf, T. Coleman, L. R. Long, S. Antani, G. R. Thoma, "An architecture for streamlining the implementation of biomedical text/image databases on the web", Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, pp. 563-568, June 2004.
11. S. Gordon, G. Zimmerman, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Content analysis of uterine cervix images: initial steps towards content based indexing and retrieval of cervigrams", Proc. of SPIE Medical Imaging, Vol. 6144, pp. 1549-1556, 2006.
12. Y. Srinivasan, D. Hernes, B. Tulpule, S. Yang, J. Guo, S. Mitra, S. Yagneswaran, B. Nutter, J. Jeronimo, B. Philips, R. Long, and D. Ferris, "A probabilistic approach to segmentation and classification of neoplasia in uterine cervix images using color and geometric features", Proceedings of SPIE Medical Imaging, Vol. 5747, pp. 995-1003, 2005
13. Y. Srinivasan, B. S. Nutter, S. D. Mitra, S. Yang, B. Phillips, L. R. Long, "Challenges in automated detection of cervical intraepithelial neoplasia", Proceedings of SPIE Medical Imaging, Vol. 6514, pp. 65140F-1-11, 2007.
14. Z. Xue, S. Antani, L. R. Long, G. R. Thoma, "Comparative performance analysis of cervix ROI extraction and specular reflection removal algorithms for uterine cervix image analysis", Proceedings of SPIE Medical Imaging, Vol. 6512, pp. 65124I-1-9, 2007.
15. J. Jeronimo, R. Long, L. Neve, B. Michael, S. Antani, and M. Schiffman, "Digital tools for collecting data from cervigrams for research and training in colposcopy", Journal of Lower Genital Tract Disease, Vol. 10, No. 1, pp. 16-25, January 2006.
16. J.W. Sellors, R. Sankaranarayanan, "Colposcopy and Treatment of Cervical Intraepithelial Neoplasia - A Beginner's Manual", Edited by J.W. Sellors and R. Sankaranarayanan, Published by the International Agency for Research on Cancer, France, 2003.
17. W. Hsu, L. R. Long, S. Antani, "SPIRS: A framework for content-based image retrieval from large biomedical databases", Proceedings of 12th International Health (Medical) Informatics Congress, Medinfo, Vol. 12, No. 1, pp. 188-192, 2007.